INFORMATION THEORY. One of the most prominent features of 20th-century technology is the development and exploitation of new communication mediums. Concurrent with the growth of devices for transmitting and processing information, a unifying theory was developed and became the subject of intensive research.

This theory, known as communication theory, or, in its broader applications, information theory, is concerned with the discovery of mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store and otherwise process information.

Some of the problems treated relate to finding the best methods of utilizing various available communication systems, the best methods of separating signals from noise and the problem of setting upper bounds on what it is possible to do with a given channel. While the central results are chiefly of interest to communication engineers, some of the concepts have been adopted and found useful in such fields as psychology and linguistics.

Information is interpreted in its broadest sense to include the messages occurring in any of the standard communication mediums such as telegraphy, radio or television, the signals involved in electronic computing machines, servomechanism systems and other data-processing devices, and even the signals appearing in the nerve networks of animals and man. The signals or messages need not be meaningful in any ordinary sense. This theory, then, is quite different from classical communication engineering theory, which deals with the devices employed but not with that which is communicated.

Central Problems of Communication Theory.—The type of communication system that has been most extensively investigated is shown in fig. 1. It consists of the following:

1. An information source which produces the raw information or message" to be transmitted.
2. A transmitter which transforms or encodes this information into a form suitable for the channel. This transformed message is called the signal.
3. The channel on which the encoded information or signal is transmitted to the receiving point. During transmission the signal may be changed or distorted. In radio, for example, there often is static, and in television transmission so-called "snow." These disturbing effects are known generally as noise, and are indicated schematically in fig. 1 by the noise source.
4. The receiver, which decodes or translates the received signal back into the original message or an approximation of it.
5. The destination or intended recipient of the information.

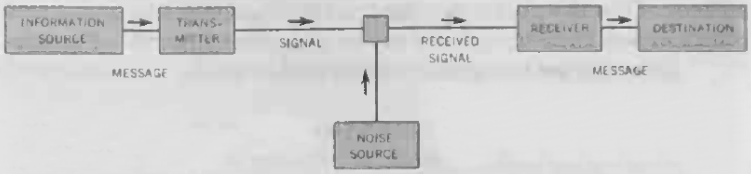

FIG. 1.—DIAGRAM OF GENERAL COMMUNICATION SYSTEM

It will be seen that this system is sufficiently general to include a wide variety of communication problems if the various elements are suitably interpreted. In radio, for example, the information source may be a person speaking into a microphone. The message is then the sound that he produces, and the transmitter is the microphone and associated electronic equipment which changes this sound into an electromagnetic wave, the signal. The channel is the space between the transmitting and receiving antennas, and any static or noise disturbing the signal corresponds to the noise source in the schematic diagram. The home radio is the receiver in this system and its sound output the recovered message. The destination is a person listening to the message.

A basic idea in communication theory is that information can be treated very much like a physical quantity such as mass or energy. A homely analogy may be drawn between the system in fig. 1 and a transportation system; for example, we can imagine an information source to be like a lumber mill producing lumber at a certain point. The channel in fig. 1 might correspond to a conveyor system for transporting the lumber to a second point. In such a situation there are two important quantities: the rate R (in cubic feet per second) at which lumber is produced at the mill, and the capacity C (in cubic feet per second) of the conveyor. These two quantities determine whether or not the conveyor system will be adequate for the lumber mill. If the rate of production R is greater than the conveyor capacity C, it will certainly be impossible to transport the full output of the mill; there will not be sufficient space available. If R is less than or equal to C, it may or may not be possible, depending on whether the lumber can be packed efficiently in the conveyor. Suppose, however, that we allow ourselves a sawmill at the source. This corresponds in our analogy to the encoder or transmitter. Then the lumber can be cut up into small pieces in such a way as to fill out the available capacity of the conveyor with 100% efficiency. Naturally in this case we should provide a carpenter shop at the receiving point to fasten the pieces back together in their original form before passing them on to the consumer.

If this analogy is sound, we should be able to set up a measure R in suitable units telling the rate at which information is produced by a given information source, and a second measure C which determines the capacity of a channel for transmitting information. Furthermore, the analogy would suggest that by a suitable coding or modulation system, the information can be transmitted over the channel if and only if the rate of production R is not greater than the capacity C. A key result of information theory is that it is indeed possible to set up measures R and C having this property.

Measurement of Information.—Before we can consider how information is to be measured it is necessary to clarify the precise meaning of "information" from the point of view of the communication engineer. Often the messages to be transmitted have meaning: they describe or relate to real or conceivable events. However, this is not always the case. In transmitting music, the meaning, if any, is much more subtle than in the case of a verbal message. In some situations the engineer is faced with transmitting a totally meaningless sequence of numbers or letters. In any

case, meaning is quite irrelevant to the problem of transmitting the information. It is as difficult to transmit a series of nonsense syllables as it is to transmit straight English text (more so, in fact). The significant aspect of information from the transmission standpoint is the fact that one particular message is chosen from a set of possible messages. What must be transmitted is a specification of the particular message which was chosen by the information source. The original message can be reconstructed at the receiving point only if such an unambiguous specification is transmitted. Thus in information theory, information is thought of as a choice of one message from a set of possible messages. Furthermore, these choices occur with certain probabilities; some messages are more frequent than others.

The simplest type of choice is a choice from two equally likely possibilities; that is, each has a probability $\frac{1}{2}$. This is the situation, for example, when a coin is tossed which is equally likely to come up heads or tails. It is convenient to use the amount of information produced by such a choice as the basic unit and this basic unit is called a binary digit or, more briefly, a "bit." The choice involved with one bit of information can be indicated schematically as in fig. 2(A). At point b either the upper or lower

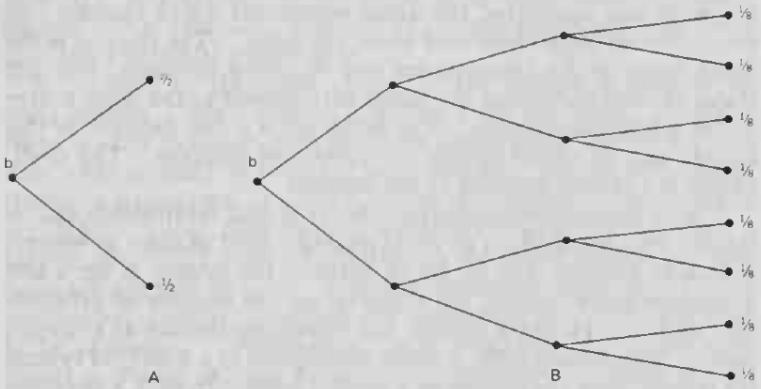

FIG. 2.—A CHOICE FROM (A) TWO POSSIBILITIES; (B) EIGHT POSSIBILITIES (*see* TEXT)

line may be chosen with probability $\frac{1}{2}$ for each possibility.

If there are N possibilities, all equally likely, the amount of information is given by $\log_2 N$. The reason for this can be seen from fig. 2(B), where there are eight possibilities each with probability $\frac{1}{8}$. The choice can be imagined to occur in three stages, each involving one bit. The first bit corresponds to a choice of either the first four or the second four of the eight possibilities, the second bit corresponds to the first or second pair of the four chosen, and the final bit determines the first or second member of the pair. It will be seen that the number of bits required is $\log_2 N$, in this case $\log_2 8 = 3$.

If the probabilities are not equal, the formula is more complicated. When the choices have probabilities $p_1, p_2, \ldots, p_n$, the amount of information H is given by:

$$H = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \ldots + p_n \log_2 p_n)$$

This formula for amount of information gives values ranging from zero—when one of the two events is certain to occur (*i.e.*, has a probability of 1) and all others are certain not to occur (probability 0)—to a maximum value of $\log_2 N$ when all events are equally probable (probability $1/N$). These situations correspond intuitively to the minimum information produced by a particular event (when it is already certain what will occur) and the greatest information or the greatest prior uncertainty of the event.

The parlor game "Twenty Questions" illustrates some of these ideas. In this game, one person thinks of an object and the other players attempt to determine what it is by asking not more than 20 questions that can be answered "yes" or "no." According to information theory each question can, by its answer, yield anywhere from no information to $\log_2 2$ or one bit of information, depending upon whether the probabilities of "yes" and "no" answers are very unequal or approximately equal. To obtain the greatest amount of information, the players should ask questions that subdivide the set of possible objects, as nearly as possible, into two equally likely groups. For example, if they have established by previous questions that the object is a town in the United States, a good question would be, "Is it east of the Mississippi?" This divides the possible towns into two roughly equal sets. The next question then might be, "Is it north of the Mason-Dixon line?" If it were possible to choose questions which always had the effect of subdividing into two equal groups, it would be possible to isolate, in 20 questions, one object from approximately 1,000,000 possibilities. This corresponds to 20 bits.

The formula for the amount of information is identical in form with equations representing entropy in statistical mechanics, and suggests that there may be deep-lying connections between thermodynamics and information theory. Some scientists believe that a proper statement of the second law of thermodynamics requires a term relating to information. These connections with physics, however, do not have to be considered in the engineering and other applications of information theory.

Most information sources produce a message which consists not of a single choice but of a sequence of choices; for example, the letters of printed text or the elementary words or sounds of speech. The writing of English sentences can be thought of as a process of choice: choosing a first word from possible first words with various probabilities; then choosing a second, with probabilities depending on the first; and so on. This kind of statistical process is called a stochastic process, and information sources are thought of, in information theory, as stochastic processes. A more general formula for H can be given which determines the rate at which information is produced by a stochastic process or an information source.

Printed English is a type of information source that has been studied considerably. By playing a kind of "Twenty Questions" game, suitably modified, with subjects trying to guess the next letter in an English sentence, it can be shown that the information rate of written English is not more than about one bit per letter. This is a result of the very unequal frequencies of occurrence of different letters (for example, E is very common in English while Z, Q and X are very infrequent), of pairs of letters (TH is very common and QZ very rare), and the existence of frequently recurring words, phrases and so on. This body of statistical data relating to a language is called the statistical structure of the language. If all 26 letters and the space in English had equal frequencies of occurrence (*i.e.*, each had probability $\frac{1}{27}$) and the occurrence of each letter of text was independent of previous letters, the information rate would be $\log_2 27$, or about 4.76 bits per letter. Since only one bit actually is produced, English is said to be about 80% redundant.

The redundancy of English is also exhibited by the fact that a great many letters can be deleted from a sentence without making it impossible for a reader to fill the gaps and determine the original meaning. For example, in the following sentence the vowels have been deleted:

MST PPL HV LTTL DFFCLTY N RDNG THS SNTNC.

As might easily be deduced, redundancy in a language plays an important role in the science of cryptography.

Encoding Information.—An important feature of the measure of information, H, is that it determines the saving in transmission time that is possible, by proper encoding, due to the statistics of the message source. To illustrate this, consider a model language in which there are only four letters—A, B, C and D. Suppose these letters have the probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{8}$. In a long text in this language, A will occur one-half the time, B one-quarter of the time, and C and D each one-eighth of the time. Suppose this language is to be encoded into binary digits, 0 or 1, as for example in a pulse system with two types of pulse. The most direct code is the following:

$$A = 00;\ B = 01;\ C = 10;\ D = 11$$

This code requires two binary digits per letter of message. By proper use of the statistics, a better code can be constructed as follows:

$$A = 0;\ B = 10;\ C = 110;\ D = 111$$

It is readily verified that the original message can be recovered from its encoded form. Furthermore, the number of binary digits used is smaller on the average. It will be, in fact,

$$\tfrac{1}{2}(1) + \tfrac{1}{4}(2) + \tfrac{1}{8}(3) + \tfrac{1}{8}(3) = 1\tfrac{3}{4}$$

where the first term is due to the letter A, which occurs half the time and is one binary digit long, and similarly for the others. It may be found by a simple calculation that $1\frac{3}{4}$ is just the value of H, calculated for the probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$.

The result verified for this special case holds generally—if the information rate of the message is H bits per letter, it is possible to encode it into binary digits using, on the average, only H binary digits per letter of text. There is no method of encoding which uses less than this amount.

This important result in information theory gives a direct meaning to the quantity H which measures the information rate for a source or a language. It says, in fact, that H can be interpreted as the equivalent number of binary digits when the language or source is encoded in 0 and 1 in the most efficient way. For instance, if the estimate of one bit per letter, mentioned above as the rate for printed English, is correct, then it is possible to encode printed English into binary digits using, on the average, one for each letter of text; and, furthermore, no encoding method would average less than this.

Capacity of a Channel.—Now consider the problem of defining the capacity C of a channel for transmitting information. Since the rate of production for an information source has been measured in bits per second, we would naturally like to measure C in the same units. The question then becomes, what is the maximum number of binary digits per second that can be transmitted over a given channel? In some cases the answer is simple. With a teletype channel there are 32 possible symbols. By calculating $\log_2 32$ it is found that each symbol represents 5 bits, provided the possible symbols are used with equal probability. Therefore, if we can send n symbols per second, and the noise level is not high enough to introduce any errors during transmission, we can send $5n$ bits per second.

The problem of calculating the capacity of a channel is usually more complex than this example because of disturbing noise. As an example, suppose there are two possible kinds of pulse that can be transmitted in a system, a 0 pulse and a 1 pulse. Suppose further that when 0 is transmitted it is received as 0 nine-tenths of the time, but one-tenth of the time noise causes it to be received as 1. Conversely, suppose a transmitted 1 is received as 1 nine-tenths of the time but distorted into a 0 pulse one-tenth of the time. This type of channel is called a binary symmetric channel. It and other noisy channels have definite capacities that can be calculated by appropriate formulas. In this particular case, the capacity is about 0.53 bits per pulse.

The meaning of the capacity of such a noisy channel may be roughly described as follows: It is possible to construct codes that will transmit a series of binary digits at a rate equal to the capacity. This can be done in such a way that they can be decoded at the receiving point with a very small probability of error. These codes are called error-correcting codes, and are so constructed that the type of transmission errors likely to occur in the channel can be corrected at the receiving point. Finally, it is not possible to transmit at a higher rate than the channel capacity and retain this error-correcting property.

The functioning of error-correcting codes can be likened to the ability of a person to correct a reasonable number of typographical errors in a manuscript because of his knowledge of the structure and context of the language. Much of the work in information theory centres around the theory and construction of such error-correcting codes.

Band-limited Channels.—A frequently occurring restriction on communication channels is that the signals must lie within a certain band of frequencies W cycles per second wide. A result known as the sampling theorem states that a signal of this type can be specified by giving its values at a series of equally spaced sampling points $1/2W$ seconds apart. Thus it may be said that such a function has $2W$ degrees of freedom, or dimensions, per second.

If there were no noise whatever on such a channel it would be possible to distinguish an infinite number of different amplitude levels for each sample. Consequently, in principle, an infinite number of binary digits per second could be transmitted, and the capacity C would be infinite. In practice, there is always some noise, but even so if no limitations are placed on the transmitter power P, the capacity will be infinite, since at each sample point an unlimited number of different amplitude levels may be distinguished. Only when noise is present and the transmitter power is limited in some way does the capacity C become finite. This capacity depends on the statistical structure of the noise as well as the nature of the power limitation.

The simplest type of noise is resistance noise, produced in an electrical resistor by thermal effects. This type of noise is completely specified by giving its average power N.

The simplest limitation on transmitter power is the assumption that the average power delivered by the transmitter is not greater than P. If a channel is defined by these three parameters W, P and N, the capacity C can be shown to be

$$C = W\log_2 \frac{P+N}{N} \text{ (bits per second)}$$

The implication of this formula is that it is possible, by properly choosing the signal functions, to transmit $W\log_2 \frac{P+N}{N}$ binary digits per second and to recover them at the receiving point *with as small a frequency of errors as desired*. It is not possible to transmit binary digits at any higher rate with an arbitrarily small frequency of errors.

Encoding systems in current use, pulse-code modulation and pulse-position modulation, use about four times the power predicted by the ideal formula. Unfortunately, as one attempts to approach more closely this ideal, the transmitter and receiver required become more complicated and the delays increase.

The relation

$$C = W\log(1 + \frac{P}{N})$$

can be regarded as an exchange relation between the bandwidth W and the signal-to-noise ratio P/N. Keeping the channel capacity fixed, the bandwidth can be decreased provided the signal-to-noise ratio is sufficiently increased. Conversely an increase in bandwidth allows a lower signal-to-noise ratio in the channel.

One method of exchanging bandwidth for signal-to-noise ratio is shown in fig. 3. The upper curve represents a signal function whose

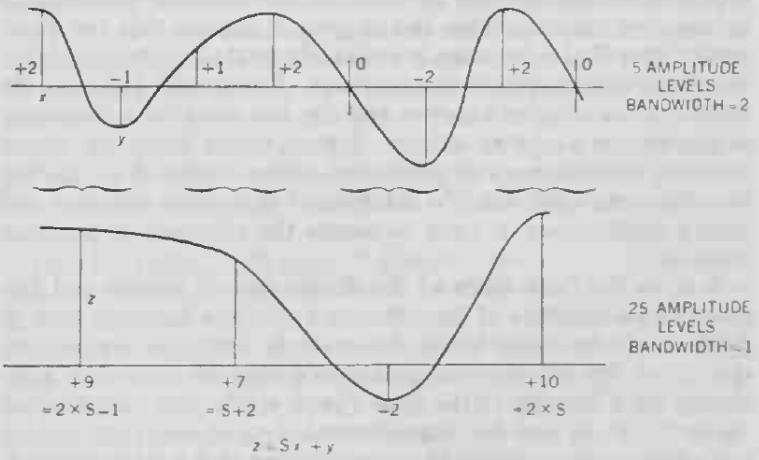

FIG. 3.—THE EXCHANGE OF BANDWIDTH FOR SIGNAL-TO-NOISE RATIO

bandwidth is such that it can be specified by giving the samples shown. Each sample has five amplitude levels. The lower curve is obtained by combining pairs of samples from the first curve as shown. If the pair of samples from the upper curve have amplitudes x and y, a single amplitude z for the lower curve is computed from the formula

$$z = 5x + y.$$

The five possible values of x combined with the five possible values of y produce 25 possible values of z which must be distin-

guished. However, the samples now occur only half as frequently; consequently the band is reduced by half, at the cost of increasing the signal-to-noise ratio. Operating this in reverse doubles the band but reduces the required signal-to-noise ratio.

To summarize, there are three essentially different ways in which bandwidth can be reduced in a system such as television or speech transmission. The first is the straightforward exchange of bandwidth for signal-to-noise ratio just discussed. The second method is utilization of the statistical correlations existing in the message. This capitalizes on particular properties of the information source feeding the channel. Finally, particular properties of the ultimate destination of the messages can be used. Thus in speech transmission the ear is relatively insensitive to phase distortion. Consequently, phase information is not as important as amplitude information, and need not be sent so accurately, resulting in a saving of bandwidth or of power. In general, the exploitation of particular "sensitivities" or "blindness" in the destination requires a proper matching of the channel to the destination.

Filtering and Prediction Problem.—Another type of problem that has been studied extensively in the field of information theory is that of determining the best devices for eliminating noise from a signal and predicting the future value of the signal. These are known as the filtering and prediction problems. The two problems may also occur in combination if it is desired to predict the future value of a noisy signal. Possible applications of the filtering problem occur in the detection of various types of communication signals which have been corrupted by noise or in the smoothing of data subject to observational error. The prediction problem, with or without filtering, may arise, for example, in weather or economic forecasting or in the control of gun directors, where it is desired to predict the future position of a moving target.

The most successful work in this general field has been carried out under the following two assumptions. First, the best prediction or filtering is interpreted to be that which minimizes the mean-square error between the computed value and the true value. Second, the devices performing these operations are assumed to perform linear operations on the signals which they filter or predict. Under these conditions, substantially complete solutions have been found specifying the characteristics of the predicting or filtering device in terms of the power spectra of the signal and of the noise.

Cryptography, Linguistics and Other Applications.—Some applications have been made in the fields of cryptography and linguistics. It is possible to formulate a theory of cryptography or secrecy systems in terms of the concepts occurring in information theory. When this is done, it appears that the information rate R of a language is intimately related to the possibility of solving cryptograms in that language. The smaller this rate, the easier such a solution becomes and the less material is necessary to render such a solution unique. Indeed, within limits, the theory becomes quantitative and predictive, giving means of calculating how much material must be intercepted in a given language and with a given cipher in order to ensure the existence of a unique solution.

A study has been made of the distribution of lengths and frequency of occurrence of the different words in a language such as English. It has been found, for example, that the relative frequency of the nth most frequent word may be expressed quite closely by a formula of the type $P(n+m)^{-b}$, with suitable constants for P, m and b. Experimental data of this type can be explained as consequences of the assumption that a language gradually evolves under continued use into an efficient communication code. (*See* CRYPTOLOGY; LINGUISTICS.)

Psychologists have discovered interesting relationships between the amount of information in a stimulus and reaction time to the stimulus. For example, an experiment can be set up in which there are four lights and four associated push buttons. The lights go on in a random order and the subject is required to press the corresponding button as quickly as possible after a light goes on. It develops that the average time required for this reaction increases linearly with an increase in the amount of information conveyed by the lights. This experimental result holds true under a wide variety of changes in the experiment: the number of lights, the probabilities of different lights, and even varying correlations between successive lights.

These results suggest that under certain conditions the human being, in manipulating information, may adopt codes and methods akin to those used in information theory.

BIBLIOGRAPHY.—C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (1949); W. Jackson (ed.), *Communication Theory* (1953); A. Feinstein, *Foundations of Information Theory* (1958); N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* (1949); C. E. Shannon, "Communication Theory of Secrecy Systems," Bell System Technical *Journal* (Oct. 1949); *Bibliography on Communication Theory,* Union Internationale des Telecommunications (1953). (CE. E. S.)